WHAT IS CLAIMED IS:

Sub A1

1. A document classification system for classifying a document based on contents of the document of which contents contains a plurality of items, said document classification system comprising:

inputting means for inputting document data corresponding to the document data;

designating means for designating at least one of the items contained in the document input by said inputting means;

converting means for converting the document data into converted data so that the converted data contains only data corresponding to the item designated by said designating means; and

classifying means for classifying the document by using the converted data produced by said converting means.

2. The document classification system as claimed in claim 1, wherein said classifying means includes document vector producing means for producing a feature vector representing a feature of the converted data so as to classify the document in accordance with the feature vector produced by said document vector producing means.

3. The document classification system as claimed in claim 1, wherein said converting means includes separation sign inserting means for inserting a predetermined sign between sets of data corresponding to the items so as to facilitate separation of each data corresponding to each item in the converted data.

4. A document classification method for classifying a document based on contents of the document of which contents contains a plurality of items, said document classification method comprising the steps of:

inputting document data corresponding to the document data;

designating at least one of the items contained in the document input in the inputting step;

converting the document data into converted data so that the converted data contains only data corresponding to the item designated in the designating step; and

classifying the document by using the converted data produced in the converting step.

5. The document classification method as claimed in claim 4, wherein the classifying step includes the step of producing a feature vector representing a feature of the converted data so as to classify the document in accordance with the feature vector.

6. The document classification system as claimed in claim 4, wherein the converting step includes the step of inserting a predetermined sign between sets of data corresponding to the items so as to facilitate separation of each data corresponding to each item in the converted data.

Sub A17

7. A processor readable medium storing program code causing a computer to classify a document based on contents of the document of which contents contains a plurality of items, comprising:

first program code means for inputting document data corresponding to the document data;

second program code means for designating at least one of the items contained in the document;

third program code means for converting the document data into converted data so that the converted data contains only data corresponding to the item designated by the second program code means; and

fourth program code means for classifying the document by using the converted data produced by the third program code means.

8. The processor readable medium as claimed in claim 7, wherein the fourth program code means includes fifth program code means for producing a feature vector representing a feature of the converted data so as to classify the document in accordance with the feature vector.

9. The processor readable medium as claimed in claim 7, wherein the third program code means includes sixth program code means for inserting a predetermined sign between sets of data corresponding to the items so as to facilitate separation of each data corresponding to each item in the converted data.

10. A document classification system for classifying a document according to contents of the document, said document classification system comprising:

input means for inputting document data of the document;

analyzing means for analyzing the document data so as to obtain analysis information; 3

vector producing means for producing a document feature vector with respect to the document data based on the analysis information; 4, 5

transforming function calculating means for calculating a representation transforming function used for projecting the document feature vector onto a space in which similarity between the document feature vectors is reflected; 6

vector transforming means for transforming the document feature vector by using the representation transforming function; 6

classification means for classifying the document based on similarity between the document feature vectors transformed by the vector transforming means; and 8

classification result storing means for storing a result of classification performed by the classification means. 9

11. The document classification system as claimed in claim 10, further comprising inner product calculating means for calculating an inner product between the document feature vectors, wherein said representation transforming function calculating means calculates the representation transforming function by using the inner product.

12. The document classification system as claimed in claim 11, further comprising document similarity information setting means for setting document similarity setting information including data representing an author of the document and a date of production of the document, wherein said representation transforming function calculating means calculates the representation transforming function by using the inner product and the document similarity information.

13. The document classification system as claimed in claim 10, further comprising:

vector storing means for storing the document feature vector produced by said vector producing means; and

transforming function storing means for storing the representation transforming function calculated by said representation transforming function calculating means.

14. The document classification system as claimed in claim 10, further comprising vector correcting means for correcting the document feature vector before the document feature vector is transformed by said vector transforming means, a correction being performed by processing one of the document feature vector and a feature dimension constituting the document feature vector in accordance with a rule established by characteristics of words extracted by said analyzing means.

15. The document classification system as claimed in claim 14, further comprising transforming function correcting means for correcting the representation transforming function calculated by said transforming function calculating means when the feature dimension is changed due to a correction of the document feature vector by said vector correcting means so that the document feature vector is transformed by said vector transforming means in accordance with the changed feature dimension.

16. The document classification system as claimed in claim 10, further comprising:

transforming function correction instructing means for sending an instruction regarding a process to be applied on a feature dimension of the representation transforming function; and

transforming function correcting means for correcting the representation transforming function based on a content of the instruction sent from said transforming function correction instructing means.

17. The document classification system as claimed in claim 16, wherein the process indicated in the content of the instruction is performed by using data of an arbitrary document vector.

18. The document classification system as claimed in claim 16, wherein the process indicated in the content of the instruction is performed by using the document feature vectors.

19. The document classification system as claimed in claim 16, wherein the process indicated in the content of the instruction is performed by using the analysis information obtained by said analyzing means.

20. The document classification system as claimed in claim 16, wherein the process indicated in the content of the instruction is performed by using the result of classification stored in said classification-result storing means.

21. The document classification system as claimed in claim 10, further comprising:

an initial cluster centroid designating means for designating an initial cluster centroid; and

initial cluster centroid registering means for registering the initial cluster centroid designated by said initial cluster centroid designating means,

wherein said classification means classifies the document in accordance with the initial cluster centroid registered by said initial cluster centroid registering means.

22. The document classification system as claimed in claim 21, wherein the initial cluster

centroid designated by said initial cluster centroid designating means is arbitrary document vector data.

23. The document classification system as claimed in claim 21, wherein the initial cluster centroid designated by said initial cluster centroid designating means is the document feature vector.

24. The document classification system as claimed in claim 21, wherein the initial cluster centroid designated by said initial cluster centroid designating means is the analysis information obtained by said analyzing means.

25. The document classification system as claimed in claim 21, wherein the initial cluster

centroid designated by said initial cluster centroid designating means is the result of classification stored by said classification-result storing means.

26. A document classification method for classifying a document according to contents of the document, said document classification method comprising the steps of:

inputting document data of the document;

analyzing the document data so as to obtain analysis information;

producing a document feature vector with respect to the document data based on the analysis information;

calculating a representation transforming function used for projecting the document feature vector onto a space in which similarity between the document feature vectors is reflected;

transforming the document feature vector by using the representation transforming function;

classifying the document based on similarity between the document feature vectors transformed in the step of transforming; and

storing a result of classification performed in the step of classifying.

27. The document classification method as claimed in claim 26, further comprising the step of calculating an inner product between the document feature vectors, wherein the representation transforming function is calculated by using the inner product.

28. The document classification method as claimed in claim 27, further comprising the step of setting document similarity setting information including data representing an author of the document and a date of production of the document, wherein the representation transforming function is calculated by using the inner product and the document similarity information.

29. The document classification method as claimed in claim 26, further comprising the steps of:

storing the document feature vector produced in the step of producing said document feature vector; and

storing the representation transforming function calculated in the step of calculating said representation transforming function.

30. The document classification method as claimed in claim 26, further comprising the step of correcting the document feature vector before the document feature vector is transformed in the step of transforming, a correction being performed by processing one of the document feature vector and a feature dimension constituting the document feature vector in accordance with a rule established by characteristics of words extracted in the step of analyzing.

31. The document classification method as claimed in claim 30, further comprising the step of correcting the representation transforming function calculated in the step of calculating when the feature dimension is changed due to a correction of the document feature vector in the step of correcting so that the document feature vector is transformed in the step of transforming in accordance with the changed feature dimension.

32. The document classification method as claimed in claim 26, further comprising the steps of:

sending an instruction regarding a process to be applied on a feature dimension of the representation transforming function; and

correcting the representation transforming function based on a content of the instruction sent in the step of sending.

33. The document classification method as claimed in claim 32, wherein the process indicated in the content of the instruction is performed by using data of an arbitrary document vector.

34. The document classification method as claimed in claim 32, wherein the process indicated in the content of the instruction is performed by using the document feature vectors.

35. The document classification method as claimed in claim 32, wherein the process indicated in the content of the instruction is performed by using the analysis information obtained by said analyzing means.

36. The document classification method as claimed in claim 32, wherein the process indicated in the content of the instruction is performed by using the result of classification stored in said classification-result storing means.

37. The document classification method as claimed in claim 26, further comprising the steps of:

designating an initial cluster centroid; and

registering the initial cluster centroid designated in the step of designating,

wherein the document is classified in accordance with the initial cluster centroid registered in the step of registering.

38. The document classification method as claimed in claim 37, wherein the initial cluster centroid designated in the step of designating is arbitrary document vector data.

39. The document classification method as claimed in claim 37, wherein the initial cluster centroid designated in the step of designating is the document feature vector.

40. The document classification method as claimed in claim 37, wherein the initial cluster centroid designated in the step of designating is the analysis information obtained in the step of analyzing.

41. The document classification method as claimed in claim 37, wherein the initial cluster centroid designated in the step of designating is the result of classification stored in the step of storing.

Sub A1

42. A processor readable medium storing program code causing a computer to classify a document according to contents of the document, comprising:

first program code means for inputting document data of the document;

second program code means for analyzing the document data so as to obtain analysis information;

third program code means for producing a document feature vector with respect to the document data based on the analysis information;

fourth program code means for calculating a representation transforming function used for projecting the document feature vector onto a space in which similarity between the document feature vectors is reflected;

fifth program code means for transforming the document feature vector by using the representation transforming function;

sixth program code means for classifying the document based on similarity between the document feature vectors transformed by the fifth program code means; and

seventh program code means for storing a result of classification performed by the classification means.

43. The processor readable medium as claimed in claim 42, further comprising eighth program code means for calculating an inner product between the document feature vectors, wherein the representation transforming function is calculated by using the inner product.

44. The processor readable medium as claimed in claim 43, further comprising ninth program code means for setting document similarity setting information including data representing an author of the document and a date of production of the document, wherein the representation transforming function is calculated by using the inner product and the document similarity information.

45. The processor readable medium as claimed in claim 42, further comprising:

tenth program code means for storing the document feature vector produced by the third program code means; and

eleventh program code means for storing the representation transforming function calculated by the fourth program code means.

46. The processor readable medium as claimed in claim 42, further comprising twelfth program code means for correcting the document feature vector before the document feature vector is transformed by the fifth program code means, a correction being performed by processing one of the document feature vector and a feature dimension constituting the document feature vector in accordance with a rule established by characteristics of words extracted by the second program code means.

47. The processor readable medium as claimed in claim 46, further comprising thirteenth program code means for correcting the representation transforming

function calculated by the fourth program code means when the feature dimension is changed due to a correction of the document feature vector by the twelfth program code means so that the document feature vector is transformed by the fifth program code means in accordance with the changed feature dimension.

48. The processor readable medium as claimed in claim 42, further comprising:

fourteenth program code means for sending an instruction regarding a process to be applied on a feature dimension of the representation transforming function; and

fifteenth program code means for correcting the representation transforming function based on a content of the instruction sent by the fourteenth program code means.

49. The processor readable medium as claimed in claim 42, further comprising:

sixteenth program code means for designating an initial cluster centroid; and

seventeenth program code means for registering the initial cluster centroid designated by the sixteenth program code means,

wherein the document is classified in accordance with the initial cluster centroid registered by the seventeenth program code means.

Add A17